YAPAY ZEKA İLKELERİ VE TEKNİKLERİ

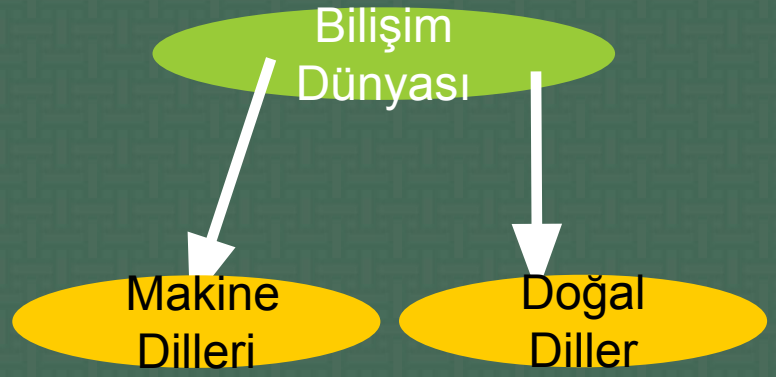
- 2020 -

DOĞAL DİL İŞLEME (DDI)
NATURAL LANGUAGE PROCESSING (NLP)

Hasan Baskın
Adli Bilişim Mühendisi
SAÜ Siber Güvenlik

Doğal Dil İşleme (DDi)
Natural Language Processing (NLP)

- Bilgisayar dünyasında iki farklı dil vardır. Bunlar Makine Dilleri dediğimiz programlama dilleri ve Doğal Diller dediğimiz insanların konuştuğu dillerdir
- Doğal Dil İşleme (DDİ), makinelerin hesaplamalı yöntemler kullanarak doğal dilde oluşturulan bir metin veya konuşmadan anlam çıkarmayı ya da konuşma veya metin sentezlemeyi hedefleyen bir yapay zeka koludur.
- Doğal dil işleme ile doğal dillerin kurallı yapısının çözümlenmesi veya yeniden üretilmesi amaçlanmıştır.
- DDİ alanında yapılan bir çalışma sadece üzerinde çalışılan dil için geçerli birçok özelliği barındırdığından, başka bir dile doğrudan aktarılamayabilir.
- DDİ alanındaki pek çok çalışma İngilizce alanında yapıldığı için, bu çalışmaların sonuçlarının veya çalışmalarda ortaya çıkan araçların Türkçe'ye doğrudan aktarılması mümkün olmamaktadır.



Dil:

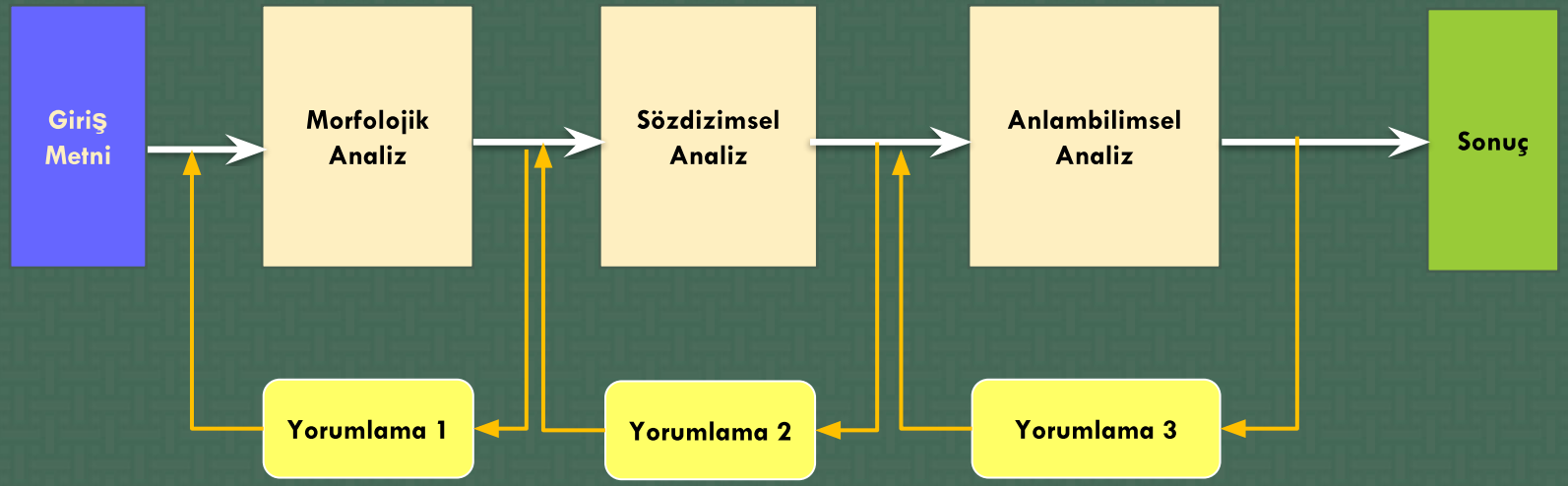
- *"Sözcük ve cümle birimleri aracılığıyla, düşünceyi konuşmayla ilişkilendiren çok seviyeli bir sistemdir"*

Noam Chomsky

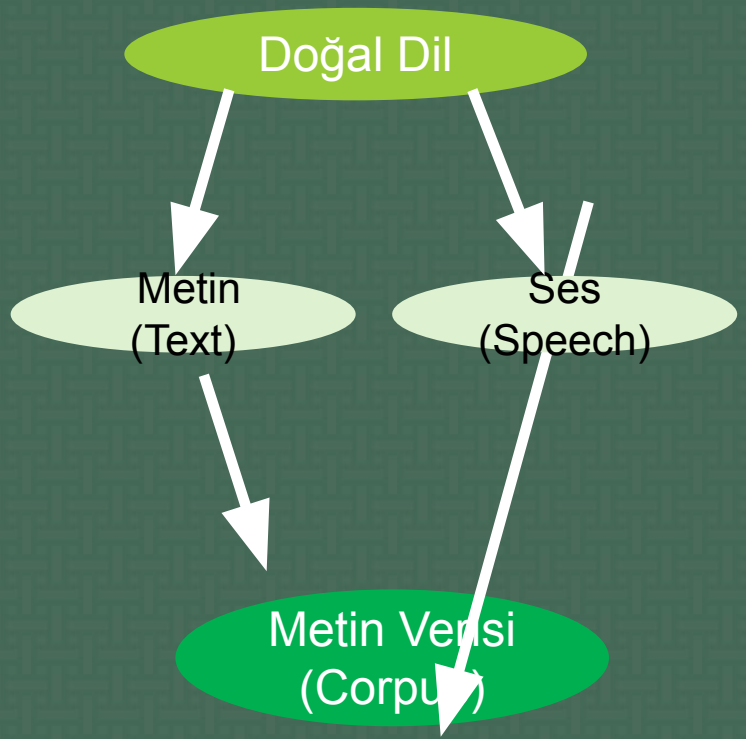
(Dilbilimci, Bilişsel Bilimci)

Doğal Dil İşleme (DDi)

- Metin ve konuşmalar Doğal Dil ile üretilir.
- Metin (Text), Yazılı verileri doğal dil işleme modüllerine uygun hale getirilip makine tarafından doğrudan işlenip analiz yapılır.
- Ses (Speech), Veriler üzerinde doğrudan işlem yapılamıyor, öncelikle ses verileri yazılı metinlere çevrilir.

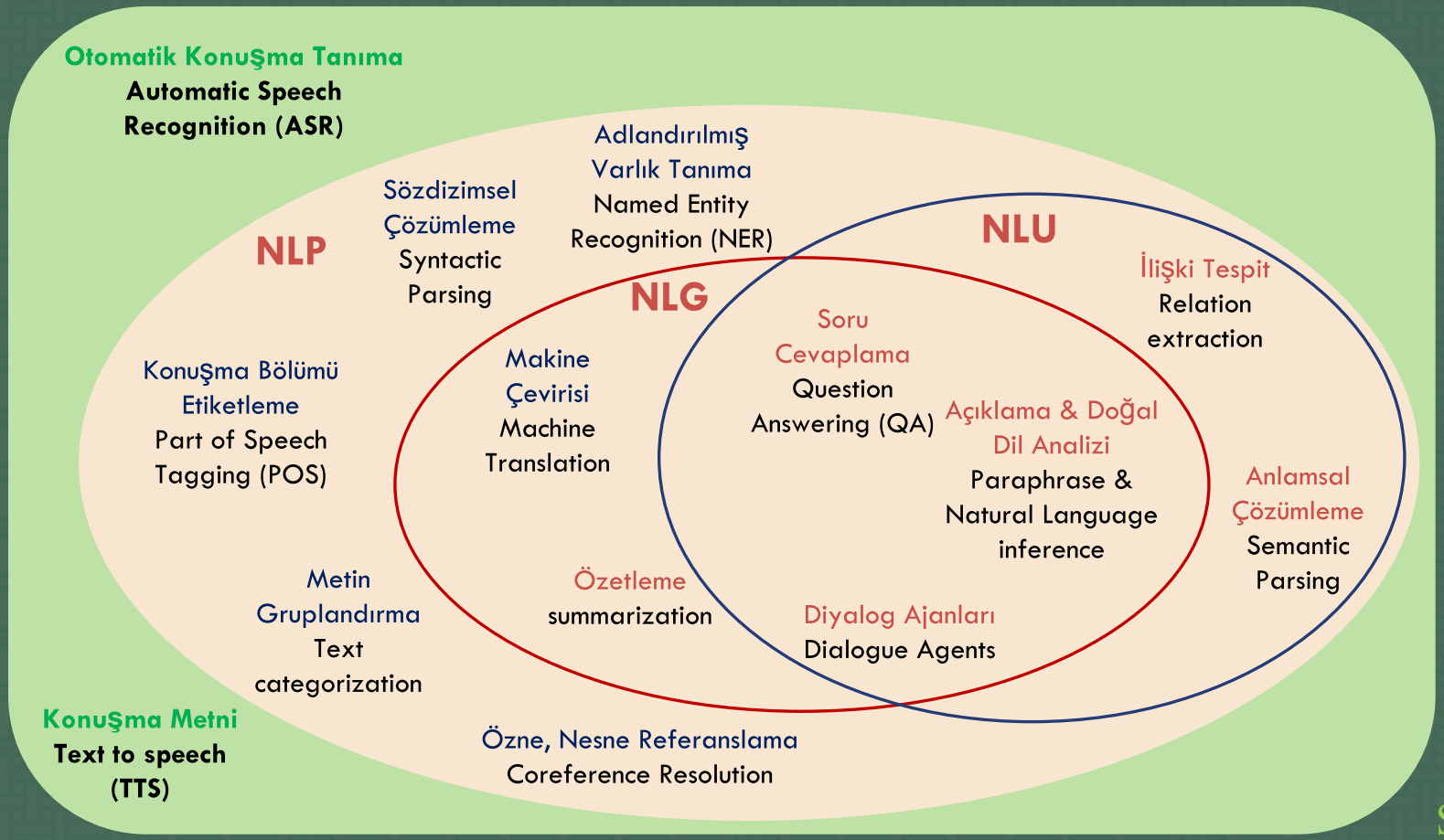


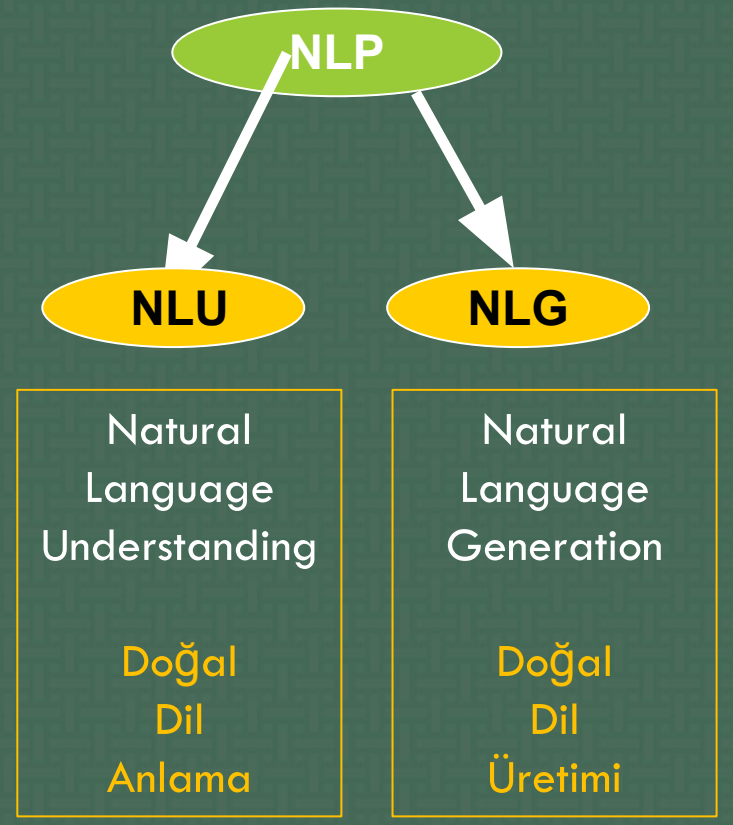
Şekil .. Dilin matematiksel analizi.



Şekil … Derlem oluşum

Natural Language Processing (NLP)





Şekil … Terminology: NLP, NLU, NLG

Doğal Dil Anlama (DDA)
Natural Language Understanding (NLU)

- Doğal dil işleme terminolojisinin bir alt kümesi olan NLU (Natural Language Understanding), yani Doğal Dili Anlama zor bir sorun olarak kabul edilmektedir.
- NLU, yazılı metnin arkasındaki anlamı anlamak için girdi verilerini yapılandırmaya çalışır.
- Makinelere insan gibi davranarak önüne gelen bir metni yorumlama yeteneği kazandırılmaya çalışılmaktadır.
- Örneğin, konuşma tanıma yazılımı konuşmayı metne dönüştürdükten sonra, NLU sistemi anlamını çözmek için bu metni veri olarak alıyor.

- Girdi metne ait farklı kelimelerin aynı veya farklı anlama sahip olması, kelimeler arasındaki ilişkilerden dolayı anlamın değişmesi, işlem yapılan dil kurallarının ve yapısının bilinmesi, NLU sistemlerinin karşılaştığı zorluklardan bazılarıdır.

Şekil DDA Doğal Dil Anlama süreçleri.

Doğal Dil İşleme (DDi)

Doğal Dilin anlaşılmasında karşılaşılan zorluklar:

- Kuralsız ve anlaşılmaz konuşmalar (sa. inŞ. Konuyu çakmışsınızdır.)
- Kuralsız ve bozuk yazılar (bbu yazdılkarımı algoritmalar düzlttebilliyor)
- Konuşmayı bölme (birden fazla cümle var duraksamadan konuşmak zor olmuyor mu?)
- Metni bölme (paragraflardan anlama, bağlama uygun cümleler oluşturmak)
- Anlam belirsizliklerini giderme (Aralıkta yapılacak toplantının yeri değişti.)

  (Aralık – ay-, aralık -boşluk-)
- Sözdizimsel belirsizlikleri giderme

  (Battal kayısıları misafirlere aç oldukları için verdi.

  Battal kayısıları misafirlere tatlı oldukları için verdi.)

Doğal Dil İşleme (DDi)

Bu çözümlemenin insana getireceği kolaylıklar,

- Yazım yanlışlarının düzeltilmesi (word processing)
- Yazılı dokümanların bir dilden diğer bir dile yarı otomatik olarak çevrilmesi
- Soru-cevap makineleri (bir veri tabanına SQL ile değil de, bir doğal dil ile sorgu yöneltme ve sistemin bunu çözümleyerek bir SQL sorgusuna çevirdikten sonra sonuçları kullanıcıya vermesi)
- Bilgisayar yardımıyla dil öğretmek,
- Çok ve tek dilli sözlüklere erişmek
- Doğal dilde cümle ve metin üretmek

- Metin özetleme

Doğal Dil Üretimi (DDÜ)
Natural Language Generation (NLG)

Makinelerin NLU'dan yapılandırılmış verileri insan tarafından anlaşılabilir dile dönüştürmeye çalıştığı adımdır. Bu nedenle NLG, NLU'nun tam tersini yapar.

Bir akıllı cihaz veya arama işlevi "duyduğu" dili anladığında, sizin de anlayacağınız şekilde sizinle konuşmak için NLG Doğal Dil Üretimi kullanır.

- Örneğin, bir arama sonucunda veriler alınıyor ve anlaşılır bir dile çevriliyor. Akıllı bir cihaza "Şu anda Malatya-Anakara yolunda durum nasıl?" diye sorulduğunda bir insan gibi cevap verebilir. "Gürün Kayseri arası kar yağışlı yer yer buzlanma olabilir." veya "Yol temiz" gibi bir şey söyleyebilir.
- NLG, sohbet robotları (chatbot) çok gelişmiş hale geldi ve ticari bir çok alanda kullanılmaya başlandı.

NLG, Gelen verilere göre, onu analiz eder ve konuşma dilinde anlatılar üretir.

NLG ile bir organizasyondaki tüm hiyerarşik düzeylerdeki insanlar için birden çok dilde anlatımlar oluşturulabilir.

NLG, öncesinde şablonlardan dinamik cümle oluşturma yaparken şimdilerde Doğal Dil Üretimi, insan benzeri metinler üreten belirli bir dizi algoritmaya dayanmaktadır.

- Markov Zinciri,
- Tekrarlayan Sinir Ağı (RNN),
- Uzun Kısa Süreli Bellek (LSTM),
- Transformatör.

Doğal Dil İşleme (DDi)
Ön İşlemler

Sözcüksel Analiz (Lexical Analiz): Kelimelerin yapısını tanımlamayı ve analiz etmeyi içerir. Bir dildeki sözcüklerin ve ifadelerin toplanması anlamına gelir. Sözcüksel analiz, bir metinin tüm yığınlarının paragraflara, cümlelere ve kelimelere ayrıştırılmasıdır.

Ses veya yazılı olarak bilgisayara girilen metin (corpus-derlem) çoğu defa yazım kurallarından yoksun ve yazım ile imla hataları ile dolu olabilmektedir. Metin eğitim verisi olarak kullanılmadan önce hepsi küçük harfe çevrilir ve noktalama işaretleri çıkarılarak aşağıdaki ön işlemler uygulanır.

1) Tokenization: Corpus'un kelimelere veya cümlelere ayrılma işlemi olarak tanımlanabilir. Metnin işlenebilmesi için kelimelere (Word2Vec) veya gerektiğinde cümlelere (Seq2Seq) ayrılması gerekmektedir.

2) Stop Words: Metni işlemeye başlamadan önce uygulanması gereken ön işlemlerden birisi de Stop Words'lerin (gereksiz kelimelerin) çıkartılmasıdır. Türkçe için Stop Words'lerin bazıları "hangi, acaba, böylece, elbette" gibi kelimelerdir.

3) Stemming: Kelimelerin köklerini alma işlemidir. Aynı kökten türeyen kelimelerin farklı kelimeler olarak algılanmaması için Stemming işlemine ihtiyaç vardır. Türkçe için Zemberek uygulaması stemming işlemini yapabilmektedir.

4) Named Entity Recognation (NER): Cümle içerisindeki kişi, organizasyon, yer isimleri ve tarihler gibi kelimeleri bulma işlemine denmektedir.

Doğal Dil İşleme Yöntemleri

Metin sınıflandırma, en genel anlamı ile eldeki bir metnin önceden belirlenen sınıflardan hangisine ya da hangilerine girdiğinin belirlenmesi demektir.

- Gereksiz (spam) e-maillerin süzülmesi,
- Metnin yazarının ya da dilinin belirlenmesi,
- Belge indeksleme,
- Sözcük anlamının

Metin sınıflandırmada kullanılan çok sayıda algoritma mevcuttur.

- Kosinüs benzerliği ve
- k-ortalama kümeleme algoritması

- Doğal dil işlemede kelimelerin vektör ile temsili en yaygın kullanılan yöntemlerden birisidir.
- Mikolov ve arkadaşları tarafından geliştirilen Word2Vec yöntemi ise yapay sinir ağlarını kullanarak kelimelerin vektör şekline dönüştürülmesini sağlamaktadır.
- Bu kelime vektörlerinin koordinatları Skip-gram modelinin eğitim amacı, bir kelimenin cümle veya metin içerisinde, o kelimeyi çevreleyen (sağındaki ve solundaki) diğer kelimeleri tahmin etmek için kelime gösterimlerini bulmaktır.

Doğal Dil İşleme (DDi)
Ön İşlemler

Doğal Dil İşleme (DDİ) kelime sınıflandırma için kullanılan bazı Algoritmalar;

- Ön işlemlerden geçen metin kelime vektörleri haline getirilip sınıflandırıcılar yardımı ile model oluşturulur. Oluşturulan model metin sınıflandırma uygulamalarında kullanılabilir.
  - Naive Bayes(NB),
  - Artificial Neural Networks (ANN),
  - K-Nearest Neighbor (KNN),
  - Logistic Regression ,
  - C4.5 classifier,
  - Multi-Layer Perceptron (MLP),
  - AdaBoost, Support Vector Machine (SVM),

Örneğin; K-means verilen bir veri seti üzerinden belirli sayıda kümeyi (k adet) gruplamak için geliştirilmiş en sade ve basit algoritmadır.

Grupların belirlenmesinde şu adımlar izlenir.

1) Kelime koordinatlarını al, grup sayısını belirle ve başlangıç grup merkezlerini (centroid) belirle.
2) Her kelimeyi en uygun gruba ata ve her atama işleminden sonra atama yapılan centroid'i hesapla.
3) Yeni oluşan grubu geçmişteki grup ile kıyasla. Grupta değişim yok ise algoritmayı bitir, aksi takdirde adım 2'ye geri dön.

Doğal Dil İşleme (DDi)
Ön İşlemler

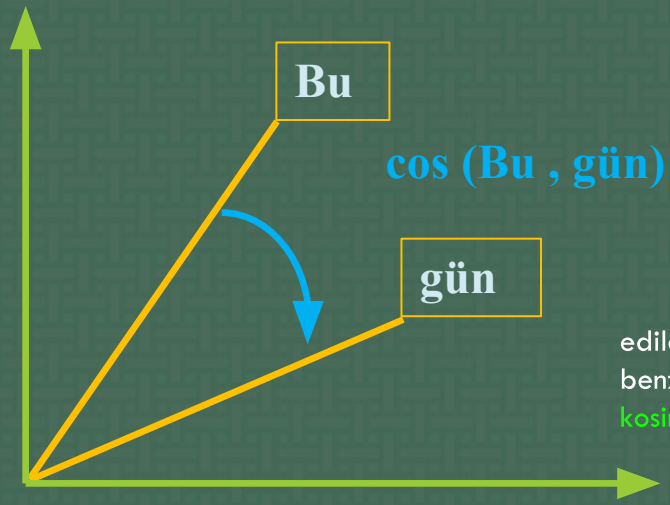
- Tokenization: Bütün bir yazıyı oluşturan her bir sözcüğü ayırma işlemidir.

Doğal dil işleme makine öğrenmesi ile elde edilen vektörlerin kelimeyi doğru olarak temsil etmelerini etkileyen üç temel etken aşağıdaki sıralanmıştır.

1-) Eğitimde Kullanılan Derlemin Büyüklüğü: Derlem büyüklüğünün artırılması, kelime vektörlerinin ağırlıkları üzerinde yapılan hata düzeltme işlemlerinin fazlalaşmasına, eğitim süresini uzamasına neden olacaktır.

2-) Eğitilen Vektörlerin Boyutu: Word2vec'te vektör boyutu 300 ile 1000 arasında olacak şekilde belirlenmesi tavsiye edilmektedir.

3-) Komşu Kelimelerin Sayısı: Word2vec için 5 ile 10 komşuluk adetleri önerilmektedir.



$$\cos(\text{Bu}, \text{gün}) = \frac{(\text{Bu}) \cdot (\text{gün})}{(\text{Bu}) + (\text{gün})}$$

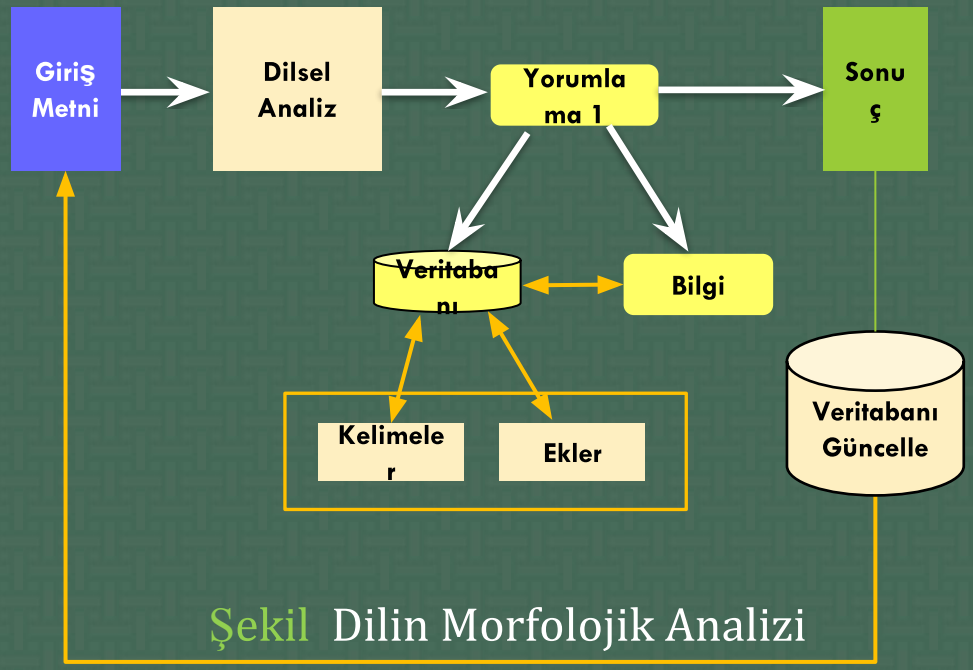
Vektör olarak temsil edilen iki kelime arasındaki benzerlik, aralarındaki açının kosinüsü olarak hesaplanabilir.

Bu	→	[0,15	0,21	0,89	0,22]
gün	→	[0,32	0,14	0,84	0,45]
hava	→	[0,28	0,75	0,34	0,68]
çok	→	[0,19	0,78	0,54	0,48]
güzel	→	[0,92	0,85	0,27	0,67]

Tablo .. Kelime vektör eşlemesi
Her bir kelime 4 boyutlu bir vektör ile tanımlanmıştır.

Doğal Dil Anlama (DDA)
Morfolojik Analiz

- Verilen cümle boşluklara ayrılarak kelimelere bölünür.
- Her bir kelime de kökleri ve aldığı eklere göre ayrılır.
- Morfolojik analiz çıktısı olarak her bir kelimenin kullanıldığı tüm durumlar gösterilir.
- Bu yüzden bir kelime için birden fazla sonuç elde edilir.
- Muğlaklık giderici ise, morfolojik analiz çıktısını işleyerek kelimenin birden fazla olan kullanımları arasından en çok kullanılan durumunu seçer.
- Kelimenin en çok kullanılan ve doğru olan kullanımı sonuçlar arasında en başta yer alır.



Şekil Dilin Morfolojik Analizi

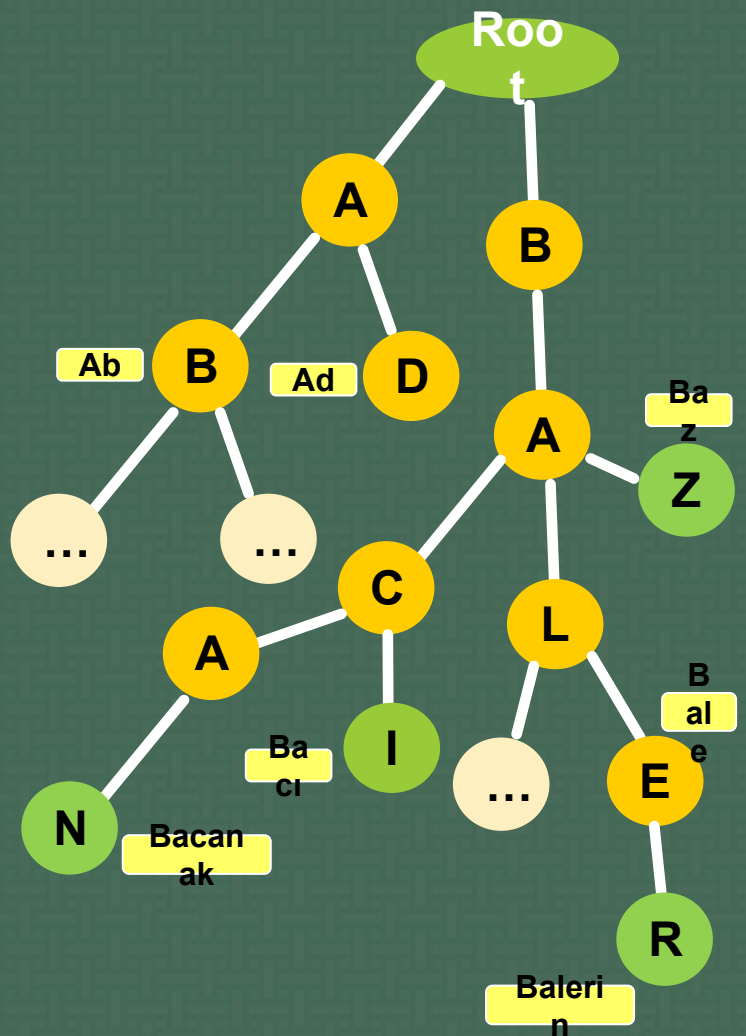
Doğal Dil İşleme (DDi)
Ön İşlemler

Normalizasyon (Normalizing): Kelimelerin köklerini alma işlemidir.

- Stem sözcükte kök veya sap anlamına gelmektedir.
- Sadece basit fiil formları ile çalışan, kelimelerden ekleri kaldıran sezgisel bir süreçtir.
- Kelimenin köküne erişebilmek için genel olarak iki yöntemden bahsedilebilir.

1- Stemming, ekli bir kelimede bulunabilen ön eklerin ve son eklerin bir listesini dikkate alarak kelimenin başlangıcını veya sonunu kesmeye çalışır. Aynı kökten türeyen kelimelerin farklı kelimeler olarak algılanmaması için Stemming işlemine ihtiyaç vardır. Türkçe için Zemberek uygulaması stemming işlemini yapabilmektedir.

2- Lemmatization: Bu metodolojinin anahtarı dilbilimidir. Doğru lemma'yı (sözlükteki bir sözcüğün kökü veya en basit biçimi) açığa çıkarmak



Şekil … Zemberek kök ağacı

Kelime Ayrıştırıcı Modül

karakter

kar, ara, rak, akt, kte, ter

kara, arak, rakt, akte, kter

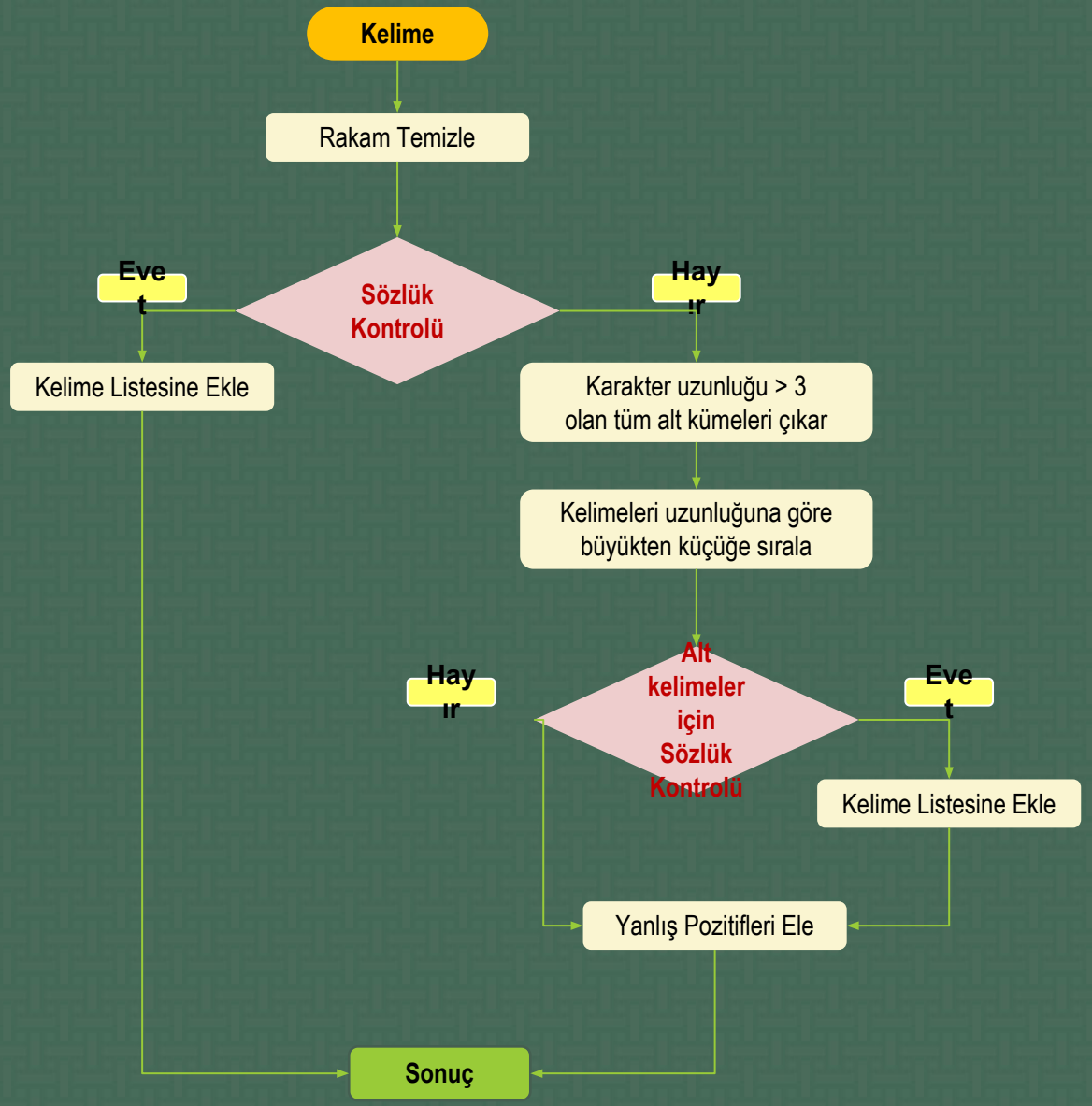
karak, arakt, rakte, akter

karakt, arakte, rakter,

karakte, arakter

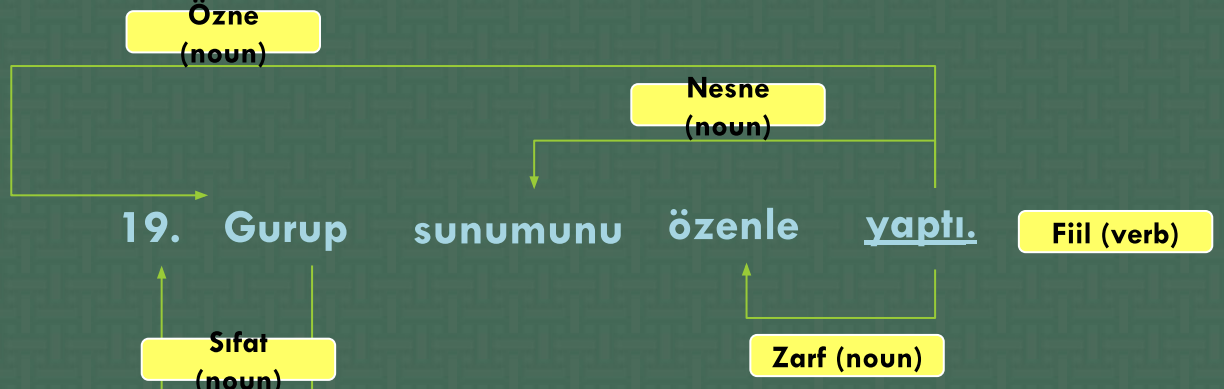
karakter

Şekil Kelime listesi oluşturma.



Syntactic Analysis (Parsing)

- Sözdizimsel Çözümleme (Ayrıştırma): Cümledeki kelimelerin analizini ve kelimeler arasındaki ilişkiyi gösterecek şekilde düzenlenmesini içerir.
- Morfolojik Analizin sonuçlarını kullanır.
- Arka arkaya gelen sözcükler yığınını, cümle birimleri olarak ifade eden bir dilbilgisi yapıya kavuşturmayı amaçlar.
- Dili anlayabilen sistemler oluşturmak için Sözdizimsel yapı ilk adım olarak gereklidir.



Şekil Dilin Sözdizimsel Analizi

Sözcüklerin Sözdizimsel Fonksiyonları:

- İsim (noun): Varlıkları tanımlar.
- Belirteç (determiner): İsmi belirler.
- Sıfat (adjective): İsmi niteler; isimle birlikte ortaya çıkar, tek başına kullanılmaz. (kırmızı top)
- Fiil (verb): Bir oluşu, bir durumu kişiye bağlayarak anlatır.
- Zarf (adverb): Fiillerin niteliğini belirtir.
- Edat (proposition): Tek başına anlam taşımaz, kendinden önceki sözcükle kullanıldığında anlamı olur. (gibi, için, kadar, karşı gibi)
- Bağlaç (conjunction): Cümleleri veya aynı görevdeki sözcükleri birbirine bağlar. (ancak, ama, fakat, ile)

Sözdizimsel Çözümleme

Sözdizim Analizinde kullanılan yöntemler:

- Yukarıdan Aşağıya Ayrıştırma
- Aşağıdan yukarıya Ayrıştırma

Word Kelime	Lemma Kök	Tag Etiket
Mavi	Mavi	+sıfat
yunus	yunus	+isim
balığı	balık	+isim, zamir
havuzdaki	havuz	+isim, edat(da, ki)
küçük	küçük	+sıfat
korkmuş	kork	+zarf, fiil, misli_gecmis
çocuğu	çocuk	+isim
kurtardı	kurtar	+fiil, dili_gecmis

Şekil Dilin Sözdizimsel Analizi

Şekil Doğal Dil İşleme Piramidi

Doğal Dil İşleme (DDi)
Natural Language Process (NLP)

A. Türkçe Karakter Dönüştürücü

- Türkçe karakterler içeren metinleri Türkçe karakterlerden arındırmak veya Türkçe karakterlerden arındırılmış bir metnin Türkçe karakterlerini geri kazandırmak için kullanılabilir. Bu araç, bir metni Türkçe karakterlerden arındırmak için oldukça basit bir yol takip etmektedir ve her bir Türkçe karakteri karşılık gelen Latince haline dönüştürmektedir.
  - Örneğin, ç harfleri c'ye çevrilirken, ı harfi i harfine çevrilmektedir.

- Bir metne Türkçe karakterleri geri kazandırmak için ise araç iki farklı yöntem içermektedir. Bu yöntemlerin ilki basit geri dönüştürücüdür (simple deasciifier). Bu yöntem, bir kelimenin var olabilecek bütün Türkçe karakter karşılıklarını oluşturmakta ve bu seçeneklerden biçimbirimsel olarak çözümlenebilen bir tanesini rassal olarak seçip sonuç olarak sunmaktadır.
  - Örneğin, cocuk girdisi için, basit geri dönüştürücü cocuk, çocuk, coçuk ve çoçuk seçeneklerini oluşturacak ve sadece çocuk biçimbirimsel olarak çözümlenebileceği için çocuk kelimesini çıktı olarak verecektir.

- Diğer yöntem ise n-karakter geri dönüştürücüdür (ngram deasciifier). Bu yöntemde ise, yine her bir kelime için ilk yöntemdeki gibi bir aday listesi oluşturulur ve ardından her bir kelimenin Türkçe'deki n-karakter olasılıkları hesaplanarak en muhtemel aday çıktı olarak verilir.

Doğal Dil İşleme (DDi)
Natural Language Processing (NLP)

B. Birimlendirici/Cümle Bölücü

- Serbest metnin birimlerini ve/veya cümlelerini saptamak için kullanılabilir.
- Bu bileşen, kural tabanlı bir bileşen olup girdiyi önceden belirlenmiş bir kural kümesini takip ederek cümlelere ve birimlerine ayırır.
- Bu kural kümesi, bir sonraki karakterin küçük/büyük harf olması gibi cümle düzeyinde kurallar içerdiği gibi, bir girdinin Türkçe'deki yaygın kısaltmalar arasında olup olmadığını kontrol etmek gibi dil düzeyinde kurallar da içerir.
- Bir girdi olarak serbest metin alır ve çıktı olarak birimlerine ayrılmış bir cümle kümesi verir.

Doğal Dil İşleme (DDi)
Natural Language Processing (NLP)

C. Yazım Denetleyici

- Verilen metindeki yazım hatalarını bulup düzelten bileşenidir.
- Her kelime için hatayı tespit edip olası doğru adaylar arasından seçim yapar.

İki farklı yazım denetleyici içermektedir.

- Basit yazım denetleyici (simple spell checker) , basit geri dönüştürücü ile benzer bir yöntem kullanır. Girdideki her kelime için her karakter gezilip bu karakter olası bütün karakterlerle değiştirilerek mümkün olabilecek bütün kelimeler oluşturulur ve bunlardan biçimbilimsel olarak çözümlenebilenlerden bir tanesi rassal olarak seçilir.

- N-karakter yazım denetleyici (ngram spell checker), benzer şekilde n-karakter geri dönüştürücü ile aynı mantığı kullanmaktadır. Önce, basit yazım denetleyicide olduğu gibi kelimeler için aday listeleri hazırlanır. Daha sonra ise n-karakter modelinden bu adaylar için olasılıklar hesaplanarak, her kelime için olasılığı en yüksek olan aday çıktı olarak verilir.

Doğal Dil İşleme (DDi)
Natural Language Processing (NLP)

D. Biçimbilimsel Çözümleyici/Belirsizlik Giderici

- Verilen girdinin çözümlemesini yapmakta, sonrasında ise bu çözümlemedeki belirsizlikleri gidermektedir. Örneğin, "Yarın doktora gidecekler." cümlesinde biçimbilimsel çözümleyici her kelime için farklı çözümlemeler bulur.
- Yarın için olası iki çözümlemeye karşılık gelen anlamlar, ertesi gün ve ikinci ya da üçüncü tekil kişinin yarısı şeklinde olabilir. Biçimbilimsel belirsizlik giderici, bu belirsizliği ortadan kaldırarak Yarın için ertesi gün anlamına gelen çözümlemeyi çıktı olarak verir.
- Türkçe dil kurallarından oluşan bir sonlu durum makinesi kullanarak verilen metini çözümler. Girdi için olabilecek belirsizlikler düşünülmeden olası bütün çözümlemeler verilir.
- Girdi olarak çözümleme listesi alır ve belirsizliği gidererek sadece doğru çözümlemeleri verir. Bu bileşen, n-karakter modeller kullanmaktadır. Kelimeler ve çözümlemeleri için iki ayrı modelden faydalanmaktadır. Öncelikle, her çözümleme için kelime ve çözümlemesinin n-karakter olasılıkları hesaplanarak en iyi kök kelime seçilir. Sonrasında, çözümlemeler bu kökü içerenlere indirgenir. Son olarak da kalan bu çözümlemelerden n-karakter modeline göre en olası olanlar bulunur ve üstlerde yer alacak şekilde çıktı olarak verilir.

Anlamsal Analiz
Semantic Analysis

- Metinden veya sözlükten toparlanan kelimelerin tam anlamını veya sözlük anlamını sistem üzerinde oluşturulur. Ve metini anlamlılık açısından kontrol edilir.
- Sözdizimsel yapıların ve nesnelerin eşleştirilmesiyle yapılır. Anlambilimsel çözümleyici, "sıcak dondurma" gibi cümleleri göz ardı eder.
- Doğal dili, insanların yapacağı gibi yorumlamaya çalışır.
- Varlıkları (Örneğin, insanlar, kuruluşlar, yerler vb.) Tanımlayan çeşitli verileri ve hedef alandaki belirli kavramları ve aralarındaki ilişkileri bütünleştirir.
- Açık Entegrasyon (Söylem Bütünleşmesi): Cümleler arasındaki anlamların birbirlerine entegrasyonu aşamasıdır.

- Varlık İsmi Tanıma kişi, yer, organizasyon gibi önceden tanımlanmış kategorilerin metin dokümanları üzerinden çıkarılma işlemidir.

Yer = 3	Kurum = 2	Kişi = 1

Malatyaspor, Bursaspor maçı öncesi ev sahibi takımın başkanı Adil Gevrek, takımını ziyaret ettikten sonra çıkışta yaptığı basın açıklamasında morallerin yüksek olduğunu söyledi.

Malatya caddelerinde taraftarlar gün boyu yaptıkları tezahüratlarla takımlarına destek verdiler.

Rakip takımın uçakla Bursa'dan Ankara aktarmalı olarak Şehre geleceği açıklandı.

Tablo .. Varlık İsmi Tanıma

Pragmatics and Discourse Analysis
Edimbilim ve Söylem Analizi

- Söylem - Discourse: Sözdizimi ve Anlam Bilimi tümce bazında çalışır. Söylem Analizi ise birden çok cümle üzerinde çalışır. Sözcük ve cümleleri kullanıldıkları bağlam içerisinde değerlendirir.

  Sözcükler Cümleler Paragraflar Dokümanlar

- Birden fazla cümleden oluşan yazılı veya sözlü söylemleri inceler
  - Cümleler arası ilişkiler çıkarılır,
  - Söylem, başlık-giriş-gelişme-sonuç kısımlarına ayırılır,

Örneğin: Battalgazi Feribot İskelesi – "Feribot hareket halindeyken araçlara binmek yasaktır."

- Anlam 1: İlçede feribot duruyorsa araçlara binilebilir.
  - Feribotun kıyıya yanaşması beklenmelidir.
- Anlam 2: Bu ilçede feribot yoksa araçlara binilmez.
- Anlam 3: Bu ilçede feribot hareket halindeyken araçlar boştur.

- Edimbilim - Pragmatics: Bir kelimenin hangi alanda ne anlama geldiğinin bilinmesi gerekir. Cümlede kullanılan bir kelime bir alan için farklı bir anlama gelebilecekken bir başka terminolojide bambaşka anlama gelebilir.

Bir yaya geçidinde iki kişi arasında aşağıdaki konuşma geçmiş olsun…

- Turgut Özal Araştırma Hastanesi nerede, biliyor musunuz?
- Evet biliyorum (der ve yürümeye devam eder)

Her iki tarafın soru ve cevap konusunda beklentileri farklıdır.

- Ev sahibi : Öğrenci misiniz?
- Kiracı : Evet. Yüksek Lisans yapıyorum.
- Ev sahibi : Allah muvaffak etsin.
- Kiracı : Bir de bağlama çalmayı öğreniyorum.
- Ev sahibi : Bu kötü

Eğer konuşmanın bağlamının ev sahibi-kiracı arasında olduğu bilinmez ise "bağlama çalmaya" kötü denmesi anlaşılmaz.

Doğal Dil İşleme (DDi)
Natural Language Processing (NLP)

Doğal dil işleme alanında çok sayıda araç ve arayüz geliştirilmiştir.

Birçok dili destekleyen açık kaynaklı ve popüler araçlar:

- Programlama arayüzü sunan
  - SpaCy
  - NLTK [2]
  - Stanford CoreNLP
- Kullanıcı arayüzü de sunan
  - Princeton WordNet

Türkçeye özgü olan araçlar:

- Programlama arayüzü sunan
  - Zemberek
- Kullanıcı arayüzü de sunan
  - ITU Türkçe Doğal Dil İşleme Yazılım Zinciri (ITU Turkish NLP Pipeline)

Doğal Dil İşleme (DDi)
Natural Language Processing (NLP)

Zemberek Temel İşlevleri:

- Yazım denetleyici (spell checker)
- Biçimbilimsel çözümleyici (morphological analyzer)
- Türkçe karakter dönüştürücü (asciifier/deasciifier)
- Birimlendirici (tokenizer)

Zemberek açık kaynak kodlu olup çeşitli doğal dil işleme bileşenlerinden oluşan bir yazılım kütüphanesidir. Kullanıcı arayüzü bulunmadığından kullanıcı kitlesi yazılım geliştiricilerden oluşmaktadır.

İTÜ Doğal Dil İşleme Yazılım Zinciri çevrimiçi bir kullanıcı arayüzü sunmasına rağmen, açık kaynak kodlu olmadığından bu araçları değiştirme ve geliştirme imkânı tanımamaktadır.

Bu platform hem bir web arayüzüne hem de bir uygulama programlama arayüzüne (API) sahip olduğundan farklı seviyelerdeki kullanıcılar bu platformdan faydalanabilir.

ITU Türkçe Doğal Dil İşleme Yazılım Zinciri Temel İşlevleri:

- Türkçe karakter dönüştürücü (asciifier/deasciifier)
- Birimlendirici/cümle bölücü (tokenizer/sentence splitter)
- Yazım denetleyici (spell checker)
- Biçimbilimsel çözümleyici (morphological analyzer)
- Belirsizlik giderici /disambiguator)
- Varlık ismi tanıma (named entity recognizer)
- Bağımlılık çözümlemesi (dependency parser)

Doğal Dil İşleme (DDi)
Natural Language Processing (NLP)

Türkçe Nlptoolkit'in birçok bileşeni teknik nedenlerden dolayı kısıtlıdır.

- Hem açık kaynak kodlu olup hem de çevrimiçi kullanıcı arayüzüne sahiptir, böylece bir yandan geliştiricilere bu araçları geliştirme ya da kendi projelerinde kullanma imkânı verirken öte yandan arayüzü ile yazılım alt yapısı olmayan kullanıcıların bu araçlardan faydalanmasına imkân tanımaktadır.

  Bileşenleri:

  - Türkçe karakter dönüştürücü (asciifier/deasciifier)
  - Birimlendirici/cümle bölücü (tokenizer/sentence splitter)
  - Yazım denetleyici (spell checker)
  - Biçimbilimsel çözümleyici (morphological analyzer)
  - Belirsizlik giderici (disambiguator)

- Bu bileşenlerden, Türkçe karakter dönüştürücü, birimlendirici/cümle bölücü ve yazım denetleyici 10.000 karakter uzunluğundaki girdilere kadar desteklese de biçimbilimsel çözümleyici/belirsizlik giderici 1.000 karaktere kadar uzunluktaki girdilere kadar cevap vermektedir.

Doğal Dil İşleme (DDi) Uygulama Alanları

- Metin Sınıflandırma ve Kategorizasyon (Text Classification and Categorization)
- Adlandırılmış Varlık Tanıma (Named Entity Recognition (NER))
- Konuşma Bölümü Etiketleme (Part-of-Speech Tagging)
- Anlamsal Ayrıştırma ve Soru Cevaplama (Semantic Parsing and Question Answering)
- Yorum Bulma (Paraphrase Detection)
- Dil Üretimi ve Çok Belgeli Özetleme (Language Generation and Multi-document Summarization)
- Dil Çeviri (Machine Translation)
- Ses Tanıma (Speech Recognition)
- Karakter Tanıma (Character Recognition)

8 best Python Natural Language Processing (NLP) libraries

- Natural Language Toolkit (NLTK) L https://www.nltk.org/
- TextBlob https://textblob.readthedocs.io/en/dev/
- CoreNLPLink: https://stanfordnlp.github.io/CoreNLP/
- GensimLink: https://github.com/RaRe-Technologies/gensim
- spaCy https://spacy.io/
- polyglotLink: https://polyglot.readthedocs.io/en/latest/index.html
- scikit–learnLink: https://scikit-learn.org/
- PatternLink: https://www.clips.uantwerpen.be/pages/pattern

Kaynaklar

- G. Şahin, "Turkish document classification based on Word2Vec and SVM classifier," in 2017 25th Signal Processing and Communications Applications Conference (SIU), 2017, pp. 1–4
- Khan Aurangzeb et al. A review of machine learning algorithms for textdocuments classification. J Adv Inform Technol 2010;1(1):4–20.
- Blanzieri E, Bryl A. A survey of learning-based techniques of email spam filtering. Tech. rep. DIT-06-056, University of Trento, Information Engineering and Computer Science Department; 2008.
- Mitchell T. Generative and discriminative classifiers: Naive Bayes and logistic regression. Manuscript available at http://www.cs.cm.edu/ ~tom/NewChapters.html; 2005.
- https://en.wikipedia.org › wiki › Natural_language_processing https://medium.com/@datamonsters/artificial-neural-networks-in-natural-languageprocessingbcf62aa9151a#targetText=Natural%20language%20generation%20has%20many,data%2C%20and%20even%20producing%20jokes. https://www.sas.com/tr_tr/insights/analytics/what-is-natural-language-processing-nlp.html

Kaynaklar

- [1] E. Riloff, “Automatically constructing a dictionary for information extraction tasks,” in Proceedings of the National Conference on Artificial Intelligence. JOHN WILEY & SONS LTD, 1993, pp. 811–811.
- [2] “Automatically generating extraction patterns from untagged text,” in Proceedings of the national conference on artificial intelligence, 1996, pp. 1044–1049.
- [3] J.-T. Kim and D. I. Moldovan, “Acquisition of linguistic patterns for knowledge-based information extraction,” Knowledge and Data Engineering, IEEE Transactions on, vol. 7, no. 5, pp. 713–724, 1995.
- [4] J. Y. Chai and A. Biermann, “The use of lexical semantics in information extraction,” in Proceedings of the ACL Workshop on Natural Language Learning, 1997.
- [5] J. Y. Chai, A. W. Biermann, and C. I. Guinn, “Two dimensional generalization in information extraction,” in Proceedings of the Sixteenth National Conference on Artificial Intelligence, 1999, pp. 431–438.
- [6] R. Basili, M. T. Pazienza, M. Vindigni, P. Bank et al., “Corpusdriven learning of event recognition rules,” in In Proceedings of Machine Learning for Information. Citeseer, 2000.
- [7] E. Agichtein and V. Ganti, “Mining reference tables for automatic text segmentation,” in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004, pp. 20–29.
- [8] D. M. Bikel, S. Miller, R. Schwartz, and R. Weischedel,“Nymble: a high-performance learning name-finder,” in Proceedings of the fifth conference on Applied natural language processing. Association for Computational Linguistics, 1997, pp. 194–201.
- [9] V. Borkar, K. Deshmukh, and S. Sarawagi, “Automatic segmentation of text into structured records,” in ACM SIGMOD Record, vol. 30, no. 2. ACM, 2001, pp. 175–186.
- [10] K. Seymore, A. McCallum, and R. Rosenfeld, “Learning hidden markov model structure for information extraction,” in AAAI- 99 Workshop on Machine Learning for Information Extraction,
- 1999, pp. 37–42.
- [11] Tang. H, Ye. J. “A Survey for Information Extraction Method” CS411

TeŞekkür

Sakarya Üniversitesi Bilgisayar ve BiliŞim Mühendisliği Ana Bilim Dalı, Siber Güvenlik Bilim Dalı 2020-2021 Tezli Yüksek Lisans programı döneminde Yapay Zeka Teknik ve İlkeleri dersi için hazırlanan “Dogal Dil İŞleme” konulu sunumda emeği geçen Bil. Müh. İsmail Güney’e teŞekkür ederim.

Hasan Baskın

Adli BiliŞim Mühendisi